rgc.cpp

June 6, 2022

Author: Takahiro Maruki

C++ program for refining GATK genotype calls (RGC).

This C++ program is for refining hard filtered genotype calls at single nucleotide polymorphism (SNP) sites by GATK.

Input file. The input is a tab-delimited text file of hard-filtered GATK genotype calls at SNP sites and can be made using VariantsToTable in GATK. The meanings of the first nine columns are: 1) scaffold (chromosome) identifier; 2) site identifier (coordinate); 3) nucleotide of the reference sequence; 4) alternative nucleotides of the variants, 5) quality score, 6) filter status, 7) alternative allele count in genotypes, 8) total number of alleles in called genotypes, 9) sum of depths over the individuals. Thereafter, four columns on the genotype call, allele depths, coverage, and genotype quality are shown for each individual.

Output file. The output file is also a tab-delimited text file. The meanings of the first nine columns are: 1) scaffold (chromosome) identifier; 2) site identifier (coordinate); 3) nucleotide of the reference sequence; 4) alternative nucleotides of the variants, 5) inferred ancestral genotype, 6) quality score, 7) alternative allele count in genotypes, 8) total number of alleles in called genotypes, 9) sum of depths over the individuals. Thereafter, the genotype call is shown for each individual in each of the columns.

Reference

If you use this program, please cite the following paper:

Maruki, T, Ozere, A, and Cristescu, M. E., Systematic identification of single nucleotide mutations from genome-wide mutation accumulation data. In prep.

Instructions

Below are specific procedures for using the program:

1. Compile the program by typing the following command:

```
g++ -o rgc rgc.cpp -lm
```

2. Run the program by typing the following command:

./ rgc -in Small_In_rgc.txt -out Small_Out_rgc.txt

- The `-in', and `-out' options specify the input file and output file name, respectively.

- The minimum and maximum sum of depths over individuals (DP) over can be specified by the `-min_dp' and `-max_dp' options, respectively. The default values for the minimum and maximum are 1 and 2,000,000,000, respectively.

- The minimum coverage required for calling individual genotypes can be specified by the `-mc' option. Its default value is 6.

- The critical values for the heterozygous and homozygous cumulative binomial probabilities in the binomial test can be specified by `-cv_het' and `-cv_hom' options, respectively. The default values for the heterozygous and homozygous cumulative binomial probabilities are 0.025 and 0.05, respectively.

- A usage help message explaining these options can be shown by typing the following command:

./rgc -h

Copyright notice

Contact

If you have difficulty using this software, please send the following information to Takahiro Maruki (takahiro.maruki@mcgill.ca):

1. Brief explanation of the problem.

2. Command used.

3. Part of the input file.

4. Part of the output file.